Indexing 30 organizes a database of protein sequences in such a way that for a given protein (represented by its feature vector 23), "similar" proteins can be found efficiently. One implementation uses the AltaVista index to index a database of proteins as represented by the generated feature vectors 23. A new "query" protein is presented to AltaVista and all similar proteins are retrieved. The similarity function used in AltaVista is modified to correspond to the vector elements of feature vectors 23. Clustering and classification techniques usually form an integral part of indexing algorithms. The main idea here is to use the index to retrieve the most similar proteins to a given query, rather than a single classification into a single structural class. This operation has important applications for biologists who are involved in drug design since a set of similar proteins can suggest multiple possible functions for a given query protein.

a4

Amendments to the specification are indicated in the attached "Marked Up Version of Amendments" (pages i - ii).

## REMARKS

The foregoing amendments to the Specification are clerical in nature. In particular, the amendments to Specification pages 6 and 7, and the addition of Table 1, merely serve to include the common nomenclature for amino acids. See MPEP §2163.07. The amendments to Specification pages 2 and 7, merely serve to clarify the text. The amendment to Specification page 11 serves to use the more commonly known name for the NI2 database, NI2 being the Applicant's internal designation for the AltaVista database.

No new matter is being introduced. Acceptance is respectfully requested.

### Information Disclosure Statement

An Information Disclosure Statement (IDS) is being filed concurrently herewith. Entry of the IDS is respectfully requested.

## CONCLUSION

In view of the above amendments and remarks, it is believed that the application is in condition for examination, and it is respectfully requested that the application be examined. If the Examiner feels that a telephone conference would expedite prosecution of this case, the Examiner is invited to call the undersigned at (978) 341-0036.

Respectfully submitted,

HAMILTON, BROOK, SMITH & REYNOLDS, P.C.

By Mary Lou Wakimura

Mary Lou Wakimura
Registration No. 31,804
Telephone ~~(781) 861-6240~~ 978 341-0036
Facsimile ~~(781) 861-9540~~ 978 341-0136

Concord, Massachusetts 01742-9133
Dated: 10/24/01

MARKED UP VERSION OF AMENDMENTS

Specification Amendments Under 37 C.F.R. § 1.121(b)(1)(iii)

**Replace the paragraph at page 2, lines 3 through 17 with the below paragraph marked up by way of bracketing and underlining to show the changes relative to the previous version of the paragraph.**

Proteins are macromolecules found in living organisms which play many roles essential to sustaining life (e.g., forming the physical framework of the organism, acting as enzymes to promote chemical reactions). A protein is composed of a sequence of several hundred amino acids. Proteins are created in living cells by translating the coding regions (genes) of the DNA sequence. Different proteins are expressed in different cells. The level of expression of different [cells] proteins determines the cell function. Since proteins are long and linear complex molecules, they "fold" to give a 3D shape. Biologists have identified four levels of structure which can influence the protein's function:

1. Primary structure--the sequence of amino acid[e]s
2. Secondary structure--the presence or absence of small "sub-folds". These are regular patterns formed by local folding of the protein (e.g., helices and sheets).
3. Tertiary structure--the final 3D shape
4. Quaternary structure--complexes formed with other proteins.

**Replace the paragraph at page 6, line 24 through page 7, line 2 with the below paragraph and Table marked up by way of bracketing and underlining to show the changes relative to the previous version of the paragraph.**

Illustrated in Fig. 1 is a computer system embodying the present invention. A digital processor 13 executes invention software program 15 in working memory. The invention software program 15 receives as input 11 a subject amino acid (i.e., protein or DNA) sequence or subsequence. The input sequence/subsequence 11 is a text string (consisting of A's, C's, T's and G's) for representing the [series of adenine, thymine, cytesine and guanine forming the molecule

corresponding to the subject amino acid sequence] sequence of amino acids. Each amino acid can be represented by one or more characters, an example of which is given in Table 1.

| Amino Acid | 3-Letter Code | 1-Letter Code |
|---|---|---|
| Alanine | Ala | A |
| Cysteine | Cys | C |
| Aspartate | Asp | D |
| Glutamate | Glu | E |
| Phenylalanine | Phe | F |
| Glycine | Gly | G |
| Histidine | His | H |
| Isoleucine | Ile | I |
| Lysine | Lys | K |
| Leucine | Leu | L |
| Methionine | Met | M |
| Asparagine | Asn | N |
| Proline | Pro | P |
| Glutamine | Gln | Q |
| Arginine | Arg | R |
| Serine | Ser | S |
| Threonine | The | T |
| Valine | Val | V |
| Tryptophan | Trp | W |
| Tyrosine | Tyr | X |

**Replace the paragraph at page 7, lines 3 through 10 with the below paragraph marked up by way of bracketing and underlining to show the changes relative to the previous version of the paragraph.**

Different amino acid sequences have different length text string representations. Hence the input sequences to invention program 15 are of varying lengths. Using a predefined set 17 of known biological fragments, the invention software program 15 performs a comparison routine 19 against the subject amino acid sequence input 11. The comparison routine 19 effectively transforms the traditional [ACTG] text representation of the subject amino acid sequence 11 into a fixed length vector 23. That is, the comparison routine 19 transforms the input sequences of varying length into respective same length (i.e., uniform length) feature vectors 23.

**Replace the paragraph at page 11, lines 2 through 13 with the below paragraph marked up by way of bracketing and underlining to show the changes relative to the previous version of the paragraph.**

Indexing 30 organizes a database of protein sequences in such a way that for a given protein (represented by its feature vector 23), "similar" proteins can be found efficiently. One implementation uses the [N12] AltaVista index to index a database of proteins as represented by the generated feature vectors 23. A new "query" protein is presented to [N12] AltaVista and all similar proteins are retrieved. The similarity function used in [N12] AltaVista is modified to correspond to the vector elements of feature vectors 23. Clustering and classification techniques usually form an integral part of indexing algorithms. The main idea here is to use the index to retrieve the most similar proteins to a given query, rather than a single classification into a single structural class. This operation has important applications for biologists who are involved in drug design since a set of similar proteins can suggest multiple possible functions for a given query protein.